DOCUMENT RESUME

ED 441 836 TM 030 871

AUTHOR Vargha, Andras; Delaney, Harold D.
TITLE Comparing Several Robust Tests of Stochastic Equality.
SPONS AGENCY Open Society Inst., New York, NY.
PUB DATE 2000-04-00
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000). Also supported by Hungarian OTKA, grants no. T018353 and T032137, and Hungarian FKFP, grant no. 0194/2000.
CONTRACT 584/1998
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Monte Carlo Methods; *Robustness (Statistics); Sample Size; Simulation; Statistical Distributions
IDENTIFIERS *Equality (Mathematics); *Stochastic Analysis; Type I Errors

ABSTRACT

In this paper, six statistical tests of stochastic equality are compared with respect to Type I error and power through a Monte Carlo simulation. In the simulation, the skewness and kurtosis levels and the extent of variance heterogeneity of the two parent distributions were varied across a wide range. The sample sizes applied were either small or moderate and equal or unequal. The tests of stochastic equality were the rank "t" test, the rank Welch test, the Fligner-Policello test, Cliff's modified Fligner-Policello test, and two variations of the last two tests that used adjusted degrees of freedom (designated FPW and FPCW). An interesting result of the study is that the two new modifications proved to be substantially more accurate with regard to their Type I error rates than the others, although they kept to a similar power level. The estimated Type I error of the FPW method at the 0.05 nominal level always fell in the range of 0.043 to 0.063 even if the variance ratio of the 2 distributions was as large as 1:16. The other ranges were: (1) 0.049-0.068 for FPCW; (2) 0.029-0.160 for the rank "t" test; (3) 0.049-0.096 for the rank Welch test; (4) 0.035-0.075 for the Fligner-Policello test; and (5) 0.040-0.078 for Cliff's test. (Contains 2 figures, 10 tables, and 22 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 836

COMPARING SEVERAL ROBUST TESTS OF STOCHASTIC EQUALITY

András Vargha
Department of General Psychology, ELTE, Hungary

Harold D. Delaney
Department of Psychology, UNM, USA

Author Notes

Much of the work reported herein resulted from the collaborative efforts of Drs. Vargha and Delaney while Vargha was a Széchenyi Professor Scholar and supported by Hungarian OTKA, grants No.: T018353 and T032157, and Hungarian FKFP, grant No.: 0194/2000. This work was also supported by the Research Support Scheme of the Open Society Support Foundation, grant No.: 584/1998.

Abstract

In the current paper six statistical tests of stochastic equality are to be compared by a Monte Carlo simulation with respect to Type I error and power. Two populations are said to be stochastically equal with respect to a variable X, if for any two independently and randomly drawn observations X_1 and X_2 from the two populations $P(X_1 > X_2) = P(X_1 < X_2)$.

In the simulation the skewness and kurtosis levels as well as the extent of variance heterogeneity of the two parent distributions were varied across a wide range. The sample sizes applied were either small or moderate, and equal or unequal. The involved tests of stochastic equality were as follows: rank t test, rank Welch test, Fligner-Policello test, Cliff's modified Fligner-Policello test as well as two modifications of the last two tests, denoted FPW and FPCW, that utilized adjusted degrees of freedom.

An interesting result obtained is that the two newly introduced test variants, FPW and FPCW, proved to be substantially more accurate with regard to their Type I error rates than the others, whereas they kept a similar power level. Specifically, the estimated Type I error of FPW at .05 nominal level always fell in the range of .043–.063 even if the variance ratio of the two distributions was as large as 1:16. The same ranges were .049–.068 for FPCW, but .029–.160 for the rank t test, .049–.096 for the rank Welch test, .035–.075 for the Fligner-Policello test, and .040–.078 for Cliff's test.

Key words: group comparison, stochastic difference, measure of stochastic superiority, stochastic equality, Fligner-Policello test, Cliff's modified Fligner-Policello test.

TM030871

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
A. Vargha
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Suppose that an experimenter wishes to know whether the scores of a variable X are of the same size in two populations or not. If X is measured on an interval scale, and one has two independent random samples from the two populations, the most commonly used statistical techniques for this purpose are Student's t test (Wilcox, 1996, p. 126), and its robust version, the Welch test (Wilcox, 1996, p. 133). With these procedures, the X-scores in the two populations are regarded to be "of the same size", if the mean of X is the same in the two populations.

However, the mean does not always characterize appropriately the level of the variable in the two populations. If the distribution of X is heavily skewed (as is Reaction Time in most cases), the extreme values "draw the mean towards them". As a result, the proportion of scores below the mean can differ greatly from the proportion above the mean. As an example, if the parent distribution is chi-square with 3 degrees of freedom, the probability that a random score will be higher than the population mean, which itself is 3, is only 39%, and accordingly the probability that a random score will be smaller than the population mean is as large as 61% (similarly, for the chi-square distribution with df = 1, 2, 4, and 5, the $P(X > \mu)$ probability equals .32, .37, .41, and .42 respectively).

The population median is a measure of location which always divides the scores of X in the population into two equal parts of 50%, provided that X is continuous (Wilcox, 1996, p. 69). However, if X is a discrete and asymmetric variable, this nice feature of the median no longer applies. As an example, if X is a five-point-scale variable, where $P(1) = .10$, $P(2) = .20$, $P(3) = .55$, $P(4) = .10$, and $P(5) = .05$, the median, M, is seemingly 3, and $P(X > M) = .10 + .05 = .15$ obviously does not equal $P(X < M) = .10 + .20 = .30$. Since the asymmetric discrete distributions play a major role in psychology and other behavioral and social sciences (see, e.g., Micceri, 1989), this greatly reduces the attractiveness of the median as a measure of location.

The trimmed mean has been proposed by several authors as an alternative location measure that is not so sensitive to the occurrence of outliers than the mean is (Wilcox, 1996, p. 15). However, the great flexibility of its definition may inevitably cause some uncertainty with regard to its interpretation. As Wilcox writes, "Currently there is no way of being certain how much trimming should be done in a given situation" (Wilcox, 1996, p. 16).

If there are a large number of competing measures for assessing the magnitude of a variable in a population, this flexibility may increase the likelihood of misunderstanding among practitioners. The conclusion that a variable, say a simple RT, has different value levels in experimental and control treatments may equally be based on a significant difference in the means, the trimmed means, or the medians. But as these location measures correspond to theoretically different conditions, their statistical tests can occasionally lead to quite different statistical results (see Wilcox, 1996, pp. 153-154).

When populations are compared with some measure of location (mean, trimmed mean, median, etc.), the equality of the populations is often presumed to be equivalent to the equality of the values of the specific location parameter in the different populations. A single representative value then is taken to indicate the location of the population in general compared to that of some other population.

A basic idea of this paper is that two populations can also be compared in a different way, by a direct comparison of the different pairs of scores $(X; Y)$, where X is any score of the dependent variable from population 1, and Y is any score from population 2. With this way of comparison population 1 will be regarded as greater than population 2, if $X > Y$ occurs more frequently than $X < Y$. Converting the occurrences into probabilities, the equality/inequality of the two populations will be determined according to the equality/inequality of the $p_+ = P(X > Y)$, $p_- = P(X < Y)$ probabilities. Referring to Agresti (1984), Hettmansperger (1984), Randles and Wolfe (1979), and Siegel and Castellan (1988), Cliff introduced the

$$\delta = P(X > Y) - P(X < Y) \tag{1}$$

difference (let us call it *stochastic difference*) for measuring the extent to which population 1 dominates population 2 with respect to the given dependent variable. He argues that "if one's primary interest is in a quantification of the statement "*X*'s tend to be higher than *Y*'s," then δ provides an unambiguous description of the extent to which this is so" (Cliff, 1993, p. 495).

To assess the difference between two continuous parent distributions McGraw and Wong used the $p_* = P(X > Y)$ probability and called it the *common language effect size indicator* or *common language statistic* (*CL*) "that is better than the available alternatives for communicating effect size to audiences untutored in statistics" (McGraw & Wong, 1992, p. 361).

For the same purpose, but for any, not necessarily continuous distribution, Vargha and Delaney (1998, 2000) introduced the *A measure of stochastic superiority*, defined as follows:

$$A_{12} = P(X > Y) + .5P(X = Y). \tag{2}$$

A_{12} is clearly a generalization of *CL*, and it is easy to see that δ and A_{12} are simple linear transformations of each other in the following way:

$$A_{12} = (\delta + 1)/2 \quad \text{and} \quad \delta = 2A_{12} - 1. \tag{3}$$

With these probability based measures, the values in populations 1 and 2 are regarded to be of the same size if

$$P(X > Y) = P(X < Y), \tag{4}$$

which occurs if and only if $\delta = 0$ or $A_{12} = .5$. If identity (4) holds, then we can conclude that neither population generally has larger values than the other. For this reason Vargha and Delaney (1998, 2000) referred to this kind of sameness of the two populations as the *stochastic equality* (denoted in the following as STE) of them, which is meaningful for any dependent variable that is at least ordinally scaled.

For the illustration of stochastic ordering we show an example. Suppose we have the following two independent samples of size three:

$$\underline{X} = (0,\ 1,\ 8) \quad \text{and} \quad \underline{Y} = (1,\ 2,\ 3).$$

For the stochastic comparison of these two samples one has to make all possible pairs of $(X; Y)$ couples. The number of different combinations is obviously 3x3 = 9, where $X > Y$ occurs in 3 cases:

$$(8;\ 1),\ (8;\ 2),\ (8;\ 3),$$

and $Y > X$ occurs in 5 cases:

$$(0;\ 1),\ (0;\ 2),\ (0;\ 3),\ (1;\ 2),\ (1;\ 3).$$

Since the proportion of $Y > X$ cases is greater than that of $X > Y$ cases:

$$Pr(Y > X) > Pr(X > Y),$$

we say: for these two samples $\underline{Y}$ *is stochastically greater than* $\underline{X}$ (or $\underline{X}$ *is stochastically smaller than* $\underline{Y}$). Analogously, if we define variable X in a population P_1 with the distribution

$$P(0) = P(1) = P(8) = 1/3,$$

and in a population P_2 with the distribution

$$P(1) = P(2) = P(3) = 1/3,$$

we obtain the same type of stochastic relation between populations P_1 and P_2 that we experienced with respect to the above two samples.

If the X dependent variable is symmetric in both P_1 and P_2 then the stochastic equality of P_1 and P_2 is equivalent to the equality of expected values ($\mu_1 = \mu_2$) and medians ($M_1 = M_2$). However, if the symmetry assumption does not hold everything can occur. Under the stochastic equality of P_1 and P_2 the equality and the inequality of the expected values can equally occur, and vice versa. The same is true with respect to the medians as well. Just in the above example we find such an astonishing situation where the mean of the X-scores ($\bar{x} = 3$) is *larger* than the mean of the Y-scores ($\bar{y} = 2$), while the sample of X-scores happens to be stochastically *smaller* than the sample of Y-scores, and the same is true with respect to the analogously derived theoretical distributions as well. Thus in the general case the concept of stochastic equality is different from that of equality of means or medians.

Now an important question: how to test the STE of two populations? In Vargha and Delaney (1998) we proved that the STE, defined by identity (4), is equivalent to another identity derived as follows. Draw two random and independent samples from populations 1 and 2, and rank them as is done in the well known Mann-Whitney-Wilcoxon test (Wilcox, 1996, pp. 365-369). In the above mentioned paper of Vargha and Delaney it was proven that the STE of two populations holds if and only if the rank scores in the two corresponding samples have identical expected values, i.e., if the expected values of the two rank-means are equal.

This equivalency reveals a way for testing STE since the equality of two expected values is of course a testable null hypothesis in classical univariate statistics, with the best known procedures being the two-sample t test (Wilcox, 1996, p. 126), and its robust version, the Welch test (Wilcox, 1996, p. 133). Thus the two-sample t test performed on the rank transforms, the *rank t test*, is a test of STE. It must be noted that the rank t test is in principle the same as the large sample version of the Mann-Whitney-Wilcoxon test (see Conover & Iman, 1981; Zimmerman & Zumbo, 1993a, 1993b; McKean & Vidmar, 1994).

In order to insure the validity of Student's two-sample t test the following three assumptions should be fulfilled: (1) independence of all individual observations from each other (this implies also the independence of the two samples), (2) normality of the parent distribution, and (3) variance homogeneity. If the original observations are independent, the first assumption holds asymptotically for the rank scores, since there is only one single constraint that makes them slightly correlate (negatively) with each other: they always sum to $N(N+1)/2$. The second and third assumption may frequently be violated in empirical studies (see, e.g., Micceri, 1989; Wilcox, 1996, p. 135), and this may invalidate the t test (Wilcox, 1996, p. 135), as well as the rank t test. Because one may not have available a good test for checking variance homogeneity, and because the Welch test generally improves upon Student's t test under the violation of normality and variance homogeneity (Wilcox,

1996, p. 135), it seems to be a reasonable choice to perform this robust alternative to t on the rank scores (rank Welch test).

For testing the $\delta = 0$ hypothesis, which is equivalent to STE defined by identity (4), Cliff (1993) mentions a robust method that was originally developed for comparing two medians due to Fligner and Policello (1981). The Fligner-Policello test procedure (denoted in the following as FP), was also suggested by Wilcox (1996, p. 369) for testing the $p_+ = P(X > Y) = .5$ hypothesis, which again is equivalent to STE in the case of continuous parent distributions. Cliff also suggests (1993, p. 499) another robust method for testing STE, which is actually a modification of the FP test (denoted in the following as FPC).

These suggested robust tests seem to make it possible to test the null hypothesis of STE without severe restrictive conditions such as identical distribution shapes or variance homogeneity. However, it is not certain that they will completely fulfill this expectation. Simulation results show that under specific conditions the Welch test may prove unacceptably poor. For example if the parent distribution is lognormal, the total sample size, $N = m + n = 100$, and the variance ratio of the two populations is 1:3, then the two-tailed Type I error rate for the Welch test at $\alpha = .05$ level can be as high as .081 for equal samples ($m = n = 50$), .101 for $m = 65$, $n = 35$, and .117 for $m = 75$, $n = 25$ (see Algina, Oshima & Lin, 1994, Table 2).

Fligner and Policello's and Cliff's tests with their approximately normal test statistics are only approximately valid. Though Fligner and Policello provide exact critical values for small samples (see Fligner & Policello, 1981, Table 1), these entries were derived under the assumption that the parent distributions have identical shapes in the two distributions. Thus it is not guaranteed theoretically that the FP test will be a valid test of STE if the distributions to be compared have different shapes (say they are oppositely skewed).

The aim of the present paper is to investigate the appropriateness (validity and power) of these suggested robust tests of STE empirically, by means of a computer simulation.

In the first section of the paper we will provide details concerning the tests used in the simulation process, suggest another two test variants for testing STE, and overview some earlier simulation studies made with some of these tests.

In section 2 we will describe the model of simulation, and provide technical details of it.

In section 3 we will discuss the obtained results of the simulation for equal and unequal sample sizes separately.

1. Details about the tests of STE

For the simulation study for testing STE we selected six statistics, whose computation formulas are summarized as follows. Let X be a variable that is least ordinally scaled, $X_1, X_2, ..., X_m$ a random sample of size m of values of X drawn from population 1 (X-sample), and $Y_1, Y_2, ..., Y_n$ a random sample of size n of values of X drawn from population 2 (Y-sample), independently of the first one. On these date samples we performed the following test procedures.

(1) Rank t test (rt)

Rank the $X_1, X_2, ..., X_m$, and the $Y_1, Y_2, ..., Y_n$ scores in the combined sample of size $N = m + n$ as in the Mann-Whitney-Wilcoxon test (see, e.g., Wilcox, 1996, p. 365). Let us denote the obtained values by $r_1, r_2, ..., r_m$ in the X-sample, and $q_1, q_2, ..., q_n$ in the Y-sample. Then compute Student's two-sample t test (see, e.g., Wilcox, 1996, p. 126) on these rank samples.

(2) Rank Welch test (rW)

Obtain the $r_1, r_2, ..., r_m$, and $q_1, q_2, ..., q_n$ rank samples the same way as with rt. Then perform the Welch test (see, e.g., Wilcox, 1996, p. 133) on these rank samples.

(3) Fligner-Policello test (FP)

For testing the equality of medians of two continuous distributions, Fligner and Policello (1981) published a modified Mann-Whitney-Wilcoxon test which did not assume the equality of population variances. The computation of the FP test is as follows. For each score X_i ($i = 1, \ldots, m$) in the X-sample determine the number of Y-scores less than X_i, and denote it by V_i. Likewise, let W_j ($j = 1, \ldots, n$) be the number of X-scores less than Y_j. The test statistic of FP is based on these V_i, and W_j values, the so called *placement scores*, and is computed as follows:

$$Z_{FP} = d/s_d. \tag{5}$$

Here in the numerator d is defined by the following formula:

$$d = (\Sigma_i V_i - \Sigma_j W_j)/(mn). \tag{6}$$

In the denominator s_d is an estimate of the *SD* of d, which, using the notations $\bar{v} = \Sigma_i V_i/m$, $\bar{w} = \Sigma_j W_j/n$, $SS_V = \Sigma_i(V_i - \bar{v})^2$, and $SS_W = \Sigma_j(W_j - \bar{w})^2$, can be written in the following form:

$$s_d = 2(SS_V + SS_W + \bar{v}\,\bar{w})^{1/2}/(mn) \tag{7}$$

(see Wilcox, 1996, p. 369). For small samples ($m, n \leq 12$) FP can be evaluated using the exact critical points reported by Fligner and Policello (1981). In other cases it can be evaluated using the normal approximation method (see Wilcox, 1996, p. 370).

In a simulation study Zumbo and Coulombe (1997) have shown that the FP test is generally not robust to the assumption of symmetric distributions if the equality of medians is tested. However, Fligner and Policello (1981, p. 164) assert that if instead of testing the equality of medians "we were interested in testing H_0: $\int GdF = .5$" (which is equivalent to $P(X_1 > X_2) = .5$ and STE; see Randles & Wolfe, 1979, p. 132), we could use the FP test without the symmetry assumption. Cliff (1993) suggests that the FP test be used for testing STE without assuming identical distributions (which implies among others things the equality of variances). The reason for this is that in the Z_{FP} test statistic the d statistic is an unbiased estimate of the $\delta = p_+ - p_-$ stochastic difference and s_d is a consistent estimator of σ_d, the *SD* of d (see Cliff, 1993, p. 499).

(4) Cliff's modified FP test (FPC)

Cliff suggested an alternative to Z_{FP} by replacing s_d with S_d, a different estimator of σ_d, which is defined by

$$(S_d)^2 = \frac{n^2\Sigma_i(d_i - d)^2 + m^2\Sigma_j(d_j - d)^2 + \Sigma_i\Sigma_j(d_{ij} - d)^2}{mn(m-1)(n-1)} \tag{8}$$

(see Cliff, 1993, identity (9)). Here $d_{ij} = sign(X_i - Y_j)$, $d_i = \Sigma_j d_{ij}/n$, $d_j = \Sigma_i d_{ij}/m$, and d, the average of all d_{ij} values, is the same as in (6) (sign(c) of any number c is defined to be –1, 0, or 1, if c is negative, zero, or positive, respectively). For evaluating FPC Cliff suggests to use the same normal approximation method which is used for FP.

These four tests were supplemented with two additional test variants.

(5) FP evaluated via Welch-like *df* (FPW)

Fligner and Policello noted that their test statistic appears to be very much like the Welch statistic computed on the placements (see Fligner & Policello, 1981, p. 164). They interpreted this

fact – just as Zumbo and Coulombe (1997, p. 140) – as a confirmation of the robustness of their test against variance heterogeneity. However, the Welch test differs from the two-sample *t* test not only in the formula of its test statistic, but also in the formula of its degrees of freedom. This reasoning leads to the idea to regard the test statistic of FP as t-distributed with a degrees of freedom computed from the placement scores the same way the degrees of freedom of the Welch test is computed from the original scores (cf. Wilcox, 1996, p. 133, formula (8.6)). Thus FPW uses the same Z_{FP} test statistic as FP, but it is evaluated via the t-distribution (regardless of the sample size) with the following approximate degrees of freedom (*df* is rounded to the nearest integer):

$$df = \frac{(a + b)^2}{a^2/(m-1) + b^2/(n-1)}, \tag{9}$$

where

$$a = SS_V/[m(m-1)] \quad \text{and} \quad b = SS_W/[n(n-1)].$$

In these formulas SS_V and SS_W are the same as in (7).

(6) FPC evaluated via Welch-like *df* (FPCW)

This test is derived from FPC the same way as FPW from FP. Therefore its test statistic equals to that of FPC, but instead of using the normal approximation method of FPC to evaluate its significance, this is evaluated via the t-distribution, with the same degrees of freedom as FPW (see formula (9) above).

So far nobody has published extensive validity results concerning statistical tests of STE for differing asymmetric distributions, where the equality of location parameters (means or medians) does not necessarily imply STE, and vice versa. Nevertheless, some evidence has been accumulated from simulation studies with rW and FP for testing the null hypothesis of means or medians involving symmetric distributions.

Zimmerman and Zumbo (1993a) showed for example, that for normal distributions "the Welch t' test protected against changes resulting from unequal variances in combination with unequal sample sizes, not only when it was performed on the initial scores, but also when it was performed on the ranks of the scores" (1993a, p. 531). Similar results, with slightly inflated Type I error rates, were obtained also for some other symmetric distributions involving the mixed normal, Cauchy, Laplace, uniform, and mixed uniform distributions (Zimmerman & Zumbo, 1992, 1993a).

Fligner and Policello (1981) reported some simulation results concerning the robustness of FP against variance heterogeneity for several symmetric continuous distributions. They found that their test "maintained its nominal level well for all situations considered" (1981, p. 167). Zumbo and Coulombe (1997) also carried out a simulation study with the FP test for testing the equality of two medians with small samples ($m, n \leq 12$), with samples drawn from either a symmetric (normal) or a heavily skewed (ex-Gaussian) parent distribution type. They found that FP performed very inconsistently for the ex-Gaussian distribution (see Table 2 in Zumbo & Coulombe, 1997), which may be due to the fact that under variance heterogeneity, equating the medians will generally not achieve STE, the proper null hypothesis of the FP test for asymmetric distributions (see Fligner & Policello, 1981, p. 164). In addition, Zumbo and Coulombe (1997) report that for the normal distribution FP performs quite conservatively even in cases where Fligner and Policello obtained

close to nominal level coverage. With our simulation study we wanted also to find an explanation to this obvious inconsistency.

Vargha and Delaney (2000) carried out a small Monte Carlo study with rt, rW, FP, and FPC. They found that if one compares identical (symmetric or asymmetric) distribution types, but allowing for differences in variances, then the latter three tests of stochastic equality (rW, FP, and FPC) prove to be substantially more robust to variance heterogeneity than rt, which is essentially the same as the Mann-Whitney-Wilcoxon test.

2. The Monte Carlo study

A Monte Carlo study was carried out to obtain some empirical information about the appropriateness of the above described six test procedures (rt, rW, FP, FPW, FPC, FPCW). With this study we wanted to obtain some evidence on whether these methods are acceptably robust to variance heterogeneity when testing the null hypothesis of STE, and to have some empirical information concerning which of them performs best. In the simulation process we systematically varied the parent distribution type (skewness-kurtosis combinations), the sample sizes and the heterogeneity of variance.

2.1 The type of the parent distribution

Random variates were generated from the generalized lambda family of distributions, which offers a variety of different shapes (Ramberg, Tadikamalla, Dudewicz, & Mykytka, 1979). These distributions are given in standardized form and can be described in terms of skewness ($\alpha_3 = \mu_3/\sigma^3$) and kurtosis ($\alpha_4 = \mu_4/\sigma^4$), where μ_3 and μ_4 are the third and fourth central moments. The generalized lambda family covers a wide range of values of skewness and kurtosis so that for any given value of skewness, several values of kurtosis can be specified (see Table 4 in Ramberg et al., 1979). For the present study three levels of skewness were applied, and for each level of skewness three levels of kurtosis were used (see Table 1). The lowest and highest levels of kurtosis always represent the most extreme levels available in Table 4 of Ramberg et al. (1979). The middle levels correspond to a medium level of kurtosis, which for a symmetric distribution gives a generalized lambda distribution having the first four moments equal to those of the standard normal. Note that the range of the possible kurtosis values depends heavily on the skewness level. At a higher skewness level both the minimal and maximal kurtosis values are higher than at a lower skewness level.

(Insert Table 1 about here)

The three levels of skewness together with the three levels of kurtosis for each yielded nine different distribution types. Crossing the distribution types of the two samples yielded 9x9 = 81 different distribution combinations, all of them included in the simulation study. The distributions listed in Table 1 all have non-negative skewness ($\alpha_3 \geq 0$). In order to investigate the appropriateness of the six tests also for oppositely skewed distribution pairs, the six asymmetric distributions appearing in Table 1 were crossed with six distributions of the same skewness levels but with an opposite sign. This yielded 6x6 = 36 more distribution pairs for the two samples.

The lambda distributions were generated in standardized form ($\mu = 0$, $\sigma = 1$) as is described in Ramberg et al. (1979). Therefore, the left-skewed distributions could be derived from the right-skewed distributions with a simple multiplication by –1. In the case of oppositely skewed distribution pairs, always the first distribution was negatively skewed.

Thus, in the simulation the total number of different distribution pairs was 81 + 36 = 117.

2.2 Sample sizes

The applied sample sizes were either small ($N = m + n = 18$) or moderate ($N = 36$), and equal ($m = n$) or unequal ($n = 2m$). Specifically, the sample sizes used were $m = n = 9$ and $m = n = 18$ in the equal sample sizes case, and $m = 6$, $n = 12$, and $m = 12$, $n = 24$ in the unequal sample sizes case.

2.3 Extent of variance heterogeneity

In the simulation the following seven *SD* ratios ($\sigma_1:\sigma_2$) of the two populations were used:

4:1, 3:1, 2:1, 1:1,1:2,1:3,1:4.

Thus, in the most extreme case the *SD* of the dependent variable in population 1 is four times as large as in population 2. Though this represents a very high level of variance heterogeneity, it can still occur in social science practice (see, e.g., Wilcox, 1996, p. 131). These *SD* ratios were created by multiplying the standardized lambda-variates with the corresponding elements of these *SD* ratios for the two samples separately. With these seven *SD* ratios, in the unequal samples unequal variances case both the direct and inverse pairing conditions could be investigated at different levels. In the simulation process all combinations of the 117 distribution pairs, 4 sample size pairs, and 7 *SD* ratios were analyzed, yielding a total number of 117x4x7 = 3276 arrangements for the two samples to be investigated.

2.4 The achievement of stochastic equality

The tested null hypothesis was STE. If the two distributions are symmetric then the equality of expected values, which obviously holds for the standardized lambda-variates, implies STE, and thus no further step has to be made. The same is true with respect to the arrangements where identical asymmetric distributions with equal variances are to be compared. To achieve STE in the asymmetric distribution and unequal variances case, the second distribution was shifted by an appropriate – positive or negative – constant. The shift constants ensuring STE were determined empirically by a Turbo Pascal program prior to the simulation process, by means of successive iterations, until STE was fulfilled. The criterion of STE was specified by the satisfaction of identity (4). A shift value was accepted if the estimated A_{12} value differed from .5 by not more than ± 0.001 three consecutive times, each time applying 400,000 randomly generated couples of variable values sampled from the two distributions. For illustration purposes the shift values for some selected distribution pairs are summarized in Table 2. Note here that the shift value is zero if and only if the two distributions are symmetric (see line 1) or are identical, implying also identical variances (see line 3, *SD* ratio = 1:1).

(Insert Table 2 about here)

2.5 The achievement of stochastic inequality

In order to assess Type I error rates of statistical tests of STE, where $A_{12} = .5$, one has to generate stochastically equal distribution pairs. However, in order to assess power rates of these tests, one has to generate stochastically unequal distribution pairs, where $A_{12} \neq .5$. Because the present simulation study focused on small and moderate sized samples, a medium and a large effect size level were employed, by setting $A_{12} = .64$, and $A_{12} = .71$. If the two parent distributions are normal (where STE is equivalent to the equality of expected values), these effect size values correspond exactly to the medium ($\Delta = .5$) and the large ($\Delta = .8$) effect size levels using Cohen's convention (see Cohen, 1977, p. 26, or Wilcox, 1996, p. 157). These levels of stochastic inequality for the 3276

different arrangements have been achieved by appropriate shift constants that have been determined with the same manner as in the STE case (see section 2.4 above).

2.6 Technical details of the simulation process

The study was conducted on a Pentium 200 MHz IBM PC compatible computer. The generalized lambda random variates were generated using the method described in Ramberg et al. (1979). In this generation process, Turbo Pascal's Random function was used to obtain pseudo-random uniform deviates. This is a linear congruential random-number generator that has turned out to be one of the most preferable in a recent study (Onghena, 1993), passing successfully ten criterion tests of randomness. For each choice of the 3276 simulation arrangements 100,000 simulation iterations were used for assessing Type I error rates, and 20,000 simulation iterations for assessing power rates. At each iteration, $N = m + n$ random variates of the desired type were generated. All of the six tests were then performed on the current set of N variates and evaluated in two-tailed form with significance level .05 and .1. Test statistics rt, rW, FPW, and FPCW were evaluated according to the usual t percentile values, based on the corresponding dfs (see section 1). In the case of FP and FPC, for small samples ($N = 18$) we used the exact critical points reported by Fligner and Policello (1981, Table 1), and for moderately large samples ($N = 36$) we used the normal approximation method (see Wilcox, 1996, p. 370). Finally, the proportion of rejections was determined. This was an estimate of the Type I error rate in the $A_{12} = .5$ case and an estimate of the power rate in the $A_{12} \neq .5$ case. With the applied number of replications the standard deviation of an empirical Type I error rate was $[\alpha(1-\alpha)/100000]^{1/2}$, which yielded .00069 if the true level was .05, and .00095 if the true level was .1. The standard deviation of an empirical power rate was always less than or equal to the quantity $[.5(1-.5)/20000]^{1/2} = .00158$ (the maximal SD of a binomial variable, $B(n, p)$, at fixed n is attained for $p = .5$).

3. Results

The appropriateness of a statistical test can be judged by evaluating at the same time both its validity (probability coverage) and efficiency. For this reason we will explain the results concerning Type I error and power rates together. First, the results concerning equal sample sizes will be presented (section 3.1). Next, we will present results with respect to unequal sample sizes. The latter require a special treatment due to the fact that the relationship between sample sizes and variances is one of the main determinants of the true Type I error rate of Student's two-sample t-test (see Scheffé, 1959, p. 353, Table 10.4.1), and this relationship can probably exert similar effects on some of our two-sample tests of STE (such as rt) as well.

3.1 Equal sample sizes

In the simulation 117 different distribution pairs were involved, as a result of a systematic variation of the skewness and kurtosis levels in the two samples (see section 2.1). Quite interestingly, the results showed that except for rt, the Type I error rates were not much influenced by either skewness or kurtosis level. Accordingly, for the sake of the easy inspection of the more important results, the obtained individual Type I error rate estimates for the 117 distribution pairs have been summarized, computing their means, minimums, and maximums. These summary statistics are presented for equal small samples ($m = n = 9$) in Table 3, and for equal moderate samples ($m = n = 18$) in Table 4. Because in the equal sample sizes case the SD ratios 1:2 and 2:1, 1:3 and 3:1, and 1:4 and 4:1 are equivalent conditions, their results were summarized too. Consequently, in the case of unequal variances the presented statistics in Table 3 and 4 are based on 2x117 = 234 different distribution pairs.

The averages of the power rate estimates corresponding to the same conditions can be seen in the case of the H_1: A_{12} = .64 alternative hypothesis in Table 5 and in the case of the H_1: A_{12} = .71 alternative hypothesis in Table 6.

(Insert Tables 3 to 6 about here)

The obtained results can be explained as follows:

1. In the case of equal and small samples, at α = .05 there are only two tests, FP and FPW, for which the Type I error rate averages never deviate from the nominal level by more than 20%, i.e., they remain between .04 and .06. The inspection of the smallest and largest individual estimates shows that this nice behavior of FP and FPW is true not only with respect to their averages, but also with respect to each individual distribution pair as well. FPW seems to be even slightly better than FP, because its Type I error estimates, which fall in the range .051–.058, deviate less from the nominal level than those of FP. The Type I error estimates of FP fall in the range .050–.060, and show a perceivable rise as the *SD* ratio increases (see Table 3, Type I error estimates at α = .05). Since in the case of $m = n = 9$ the power levels of FP and FPW are practically identical (see the upper left panel of Table 5 and Table 6), we are justified to claim that under the condition of equal and small samples FPW seems to work best out of the six tests being compared.

2. Under the same conditions but at the α = .1 level, only FP, FPW, and FPCW have Type I error rate averages that never deviate from the nominal level by more than 20%, i.e., they remain between .08 and .12. Among them FPW seems again to be the best. Since its Type I error estimates fall always in the range .095–.106, they never deviate from the nominal level by more than 6%, whereas the maximal deviation is 13% in the case of FPCW, and 19% in the case of FP (see the lower part of Table 3). Since the power levels of FP, FPW, and FPCW are practically identical (see the upper right panel of Table 5 and Table 6), we are justified to claim that under the condition of equal and small samples FPW seems to perform best.

3. In the case of equal and moderate samples ($m = n = 18$), at α = .05 there are only two tests, FPW and FPCW, for which the Type I error rate averages never deviate from the nominal level by more than 20%. The inspection of the smallest and largest individual estimates shows that for FPW and FPCW the individual Type I error estimates fall in the range .049–.053 and .051–.058, respectively (see the upper part of Table 4). Since in the case of equal sample sizes the power levels of FPW and FPCW are practically identical (see the lower left panel of Table 5 and Table 6), FPW seems again to be the best test, though FPCW performs almost as well as FPW.

4. Under the same conditions but at the α = .1 level, four tests (FP, FPW, FPC, and FPCW) have Type I error rates that never deviate from the nominal level by more than 20%. Among them the best are FPW and FPCW for which the maximal deviation is not more than .006, and .007 respectively (see the lower part of Table 4). Concerning power, the power rates of FP and FPC are generally higher than those of the other two tests by 2-3% (see the lower right panel of Table 5 and Table 6), but this is clearly due to their increased Type I error rates. For this reason in the moderate sample sizes case at α = .1 level again FPW and FPCW are the tests of choice.

From the above results it is quite obvious that under the condition of equal sample sizes the best performing test of STE is FPW, the Fligner-Policello test statistic with Welch's like degrees of freedom. While its power is of the same magnitude as that of its competitors, among 117x2x7 =1638 simulation arrangements the Type I error rates of FPW never exceeds the nominal level by more than 16%, at α =.05 level falling always in the range .049–.058, and at α = .1 level in the range .095–.106.

The performance of FP approaches that of FPW in the case of small sample sizes, when it is evaluated with the exact critical values, but the probability coverage of FP is somewhat more

sensitive to variance heterogeneity than that of FPW. However, in the case of larger samples, FP becomes slightly inflated and therefore no longer competes FPW effectively.

The behavior of FPCW is just the opposite of that of FP. In the small sample case it is often more inflated than FPW, but with larger samples it offers a real alternative to FPW.

The Type I errors of FPC never deviate dramatically from the nominal level. The maximal deviations never exceed the extent of 40% (such as .07 at $\alpha = .05$), but this performance is clearly weaker than that of the above three tests.

The performance of rW is by and large similar than that of FPC. It has somewhat better Type I error rates under the condition of variance homogeneity, but if the variances are different, rW becomes more inflated than FPC.

The rt test seemed to be quite acceptable for testing STE under the condition of variance homogeneity (just as the *t* test for comparing two means), but as was expected, it became greatly inflated when the *SD* ratio differed from 1:1.

3.2 Unequal sample sizes

The results again showed that the Type I error rates of the robust tests of STE were not much influenced by the shape of the distribution (depending on skewness and kurtosis). In this case the summary statistics (averages, minimums, and maximums) are presented for unequal small samples ($m = 6$, $n = 12$) in Table 7, and for unequal moderate samples ($m = 12$, $n = 24$) in Table 8. The averages of the power rate estimates corresponding to the same conditions can be seen in the case of the H_1: $A_{12} = .64$ alternative hypothesis in Table 9 and in the case of the H_1: $A_{12} = .71$ alternative hypothesis in Table 10.

(Insert Tables 7 to 10 about here)

The obtained results can be explained as follows:

1. In the case of small and unequal sample sizes, at $\alpha = .05$ FPW is the only test for which the Type I error rate averages never deviate from the nominal level by more than 20% for each *SD* ratio (see the top panel in Table 7). The inspection of smallest and largest individual estimates shows that the maximal deviation of the nominal value is .013 since the maximum of the individual Type I error estimates of FPW is .063. Likewise, the maximal Type I error estimate is .068, .070, .072, and .092 for FPCW, FP, FPC, and rW respectively (see the third panel of Table 7). The power level of FPW was substantially lower than that of only rt, and then only in the cases where larger sample sizes were paired with smaller variances (see Table 9). But, of course, in those situations the empirical Type I error level of rt was extremely high, exceeding the nominal level by more than 40% (see Table 7). Since in this case the power level of FPW is not markedly lower than that of the other four valid tests, we are justified to claim that under the condition of small and unequal samples FPW seems to work best.

By examining the individual distribution pairs producing Type I error estimates exceeding .06 (the number of such pairs was 30) we found the following interesting result. For 27 out of these 30 distribution pairs the second, for which the corresponding sample size was the larger one, had the lowest kurtosis level at its skewness level (see Table 1), and for the remaining 3 pairs the second had a medium level of kurtosis. In brief: if the parent distribution of the larger sample has low kurtosis, and the sample sizes and variances are inversely related, then FPW tends to be slightly inflated. At the same time the frequencies of the three skewness levels (0, 1, and 2) of the second distribution was 8, 12, and 10, that is the level of the skewness did not have a perceivable effect on the Type I error rate.

2. Under the same conditions but at the $\alpha = .1$ level, only FPW and FPCW has Type I error rate averages that never deviate from the nominal level by more than 20%. The largest deviation occurs with the 4:1 *SD* ratio. In this case FPW tends to be slightly inflated (for the individual Type I error estimates: min. = .104, max. = .132), whereas FPCW tends to be slightly conservative (min. = .067, max. = .090; see the first data column of the two lowest panels of Table 7). The situation is somewhat similar with the 1:4 *SD* ratio, but in the opposite direction. Here FPW proves to be slightly conservative (min. = .083, max. = .092), and FPCW inflated (min. = .100, max. = .110; see the last data column of the two lowest panels of Table 7). Considering also the obtained power estimates one can claim that with small and unequal samples, and at $\alpha = .1$, if the sample sizes and the variances are inversely related, then FPW is the test of choice, and in other cases (if the sample sizes and the variances are positively related or the variances are equal) FPCW.

Identifying the distribution pairs producing Type I error estimates exceeding .120 (the number of such pairs was 12) we found that for 9 out of these 12 distribution pairs the first, for which the corresponding sample size was the smaller one, was symmetric, and had low kurtosis level, and for the remaining 3 pairs the first was moderately skewed and had also a low kurtosis level. This means that the largest Type I error estimates occurred again at the lowest kurtosis levels.

3. In the case of moderate and unequal sample sizes ($m = 12$, $n = 24$), at $\alpha = .05$ two tests, FPW and FPCW can be characterized by excellent Type I error rates, whereas the probability coverage of the other four tests happens to be often rather poor (see the upper half of Table 8). Since in this case the power level of FPW and FPCW is only slightly lower than that of the other tests (see the third panel of Table 9 and Table 10), under these conditions FPW and FPCW are the winners with a slight advantage of FPCW.

4. Under the same conditions but at $\alpha = .1$ the same conclusion can be drawn: the best performing tests are always FPW and FPCW, with a slight advantage of FPCW (see the lower half of Table 8 and the fourth panel of Table 9 and Table 10).

Based on the above detailed results one can conclude that in the small and unequal sample sizes case FPW seems to be the most reliable test of STE. The reason for this is that FPW is never substantially weaker than any of the other tests, but each of the others produces occasionally a Type I error rate that is substantially higher than the nominal level. If the unequal sample sizes are larger, the picture is similar with the difference that in this case FPCW, producing occasionally a somewhat higher power rate than FPW, becomes a real alternative to FPW.

Disproving prior expectations FP did not prove resistant against variance heterogeneity even in the small sample case where it can be evaluated by "exact" critical values. As an example, with $m = 6$, $n = 12$ and $\alpha = .05$, the Type I error rates of FP are all about .065 at the 4:1, 3:1, and 2:1 *SD* ratios, but at the 1:4 and 1:3 *SD* ratios these rates drop to .037 (see the top panel of Table 7).

In all unequal sample settings the performance of FPC is always very similar to that of FP.

The probability coverage of rW is only acceptable where the larger *SD* is at most twice as large than the smaller one. If the *SD* ratio is more extreme, rW becomes unduly inflated.

The probability coverage of rt is only acceptable under the condition of variance homogeneity. If the larger *SD* is just twice as large than the smaller one, then the Type I error of rt will substantially differ from the nominal level.

Discussion

In this study the validity and efficiency of six statistical tests of stochastic equality (STE) have been compared by means of computer simulation. Contrasted to the equality of means, medians, or other location parameters, where the two populations are compared by means of two representative values, STE represents an equality that is based on the direct comparisons of the

elements of the two populations. Two populations are said to be stochastically equal with respect to a variable that is at least ordinally scaled, if for any two independently and randomly drawn X and Y observations from the two populations $P(X > Y) = P(X < Y)$.

For testing STE several robust methods have been suggested by different authors (Fligner and Policello, 1981; Cliff, 1993; Zumbo & Coulombe, 1997; Vargha & Delaney, 2000). However, the appropriateness of these procedures have not been proved either theoretically or empirically for conditions where the shapes (such as variances, skewness levels, kurtosis levels, etc.) of the two distributions differ. The present study undertook and performed this task.

The simulation varied the skewness and kurtosis levels over broad ranges and there were compared symmetric, as well as positively and negatively skewed distributions, crossed in all possible combinations. In the design of simulation we applied seven *SD* ratios with four levels of variance heterogeneity (σ_1:σ_2 = 1:1, 1:2, 1:3, 1:4), equal and unequal, and small and moderate sized samples. In addition to the four rank tests of STE proposed by earlier studies (rt, rW, FP, FPC), we introduced two other methods based on theoretical considerations (FPW and FPCW). The test statistics of FP and FPW, and similarly FPC and FPCW, are identical. They only differ in how they are evaluated. While the evaluation of FP and FPC is based either on exact critical values ($m, n \leq$ 12) or on a normal approximation (large sample case), FPW and FPCW is to be evaluated by the t-distribution with a degrees of freedom that can be determined analogously to that of the Welch test (see formula (9) in section 1).

Quite interestingly the results revealed that the newly suggested two tests, FPW and FPCW could be characterized by substantially better Type I error rates than the others, whereas their power was not much worse. FPW was clearly the best test in maintaining its Type I error rate very close to the nominal level. For example, at α = .05 significance level, the Type I error of FPW fell always in the range .043–.063 out of 2x7x117 = 1638 small sample arrangements, and in the range .043–.057 out of 2x7x117 = 1638 moderate sized sample arrangements. By contrast, for rW, FP, FPC, and FPCW, for small samples the same ranges were .049–.093, .035–.070, .040–.072, and .049–.068 respectively, and for moderate samples .050–.096, .054–.080, .061–.078, and .049–058 respectively.

This impressive result can be appreciated even more if we mention that the Welch test, which is known to be a robust method for comparing two means when the population variances differ, produces occasionally dramatically bad Type I error rates. As we have already mentioned, in a study Algina et al. (1994) showed that if the parent distribution is lognormal, the *SD* ratio is 1:3, the sample size ratio is 13:7, and the total sample size, $N = m + n$ is as large as 100, then the Type I error rate of the Welch test at α = .05 can be as large as .101, and if N = 500 then the Type I error rate is still as large as .064 (see Table 2 of Algina et al., 1994). When we carried out the simulation analysis with the six rank tests (see section 2) using the standardized lognormal as parent distribution, α = .05, and N = 20 (m = 13, n = 7), then the Type I error rate of FPW was within the .044–.061 range for each of the seven *SD* ratios applied (even for 1:4 and 4:1). Also, when the total sample size was as large as 100 (m = 65, n = 35), the Type I error rates of FPW were all within the .047–.053 range.

For moderate samples FPCW seems to be a good alternative to FPW. With regard to FPCW further simulations are needed in order to clarify the sample size level for which the Type I error rate does not exceed the nominal level by more than 20%.

In our simulation design, for the sample sizes we only used the 1:1, 1:2 ratios. Since in social sciences occasionally occur more extreme ratios, further simulations are needed to obtain empirical evidence of the appropriateness of FPW under these conditions as well.

The family of generalized lambda-distributions covers a really broad range of continuous distribution types (see Ramberg et al., 1979). However, they do not well represent the bimodal distributions, which can be generated for example by an appropriate mixture of two normals. Thus

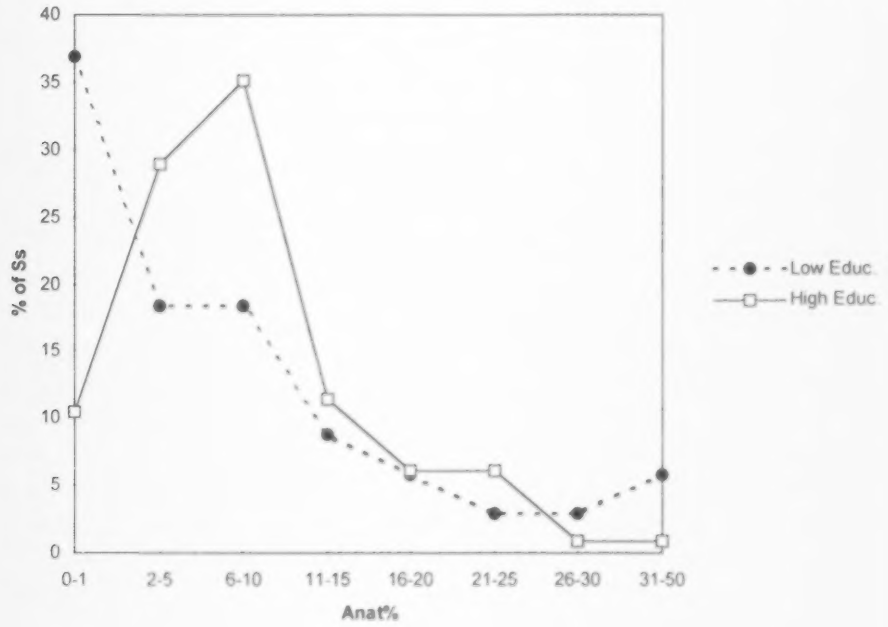
there is a good reason to extend the simulation study into this direction too, and the same can be suggested with respect to the discrete distribution types as well, which occur in social science research very frequently (see, e.g., Micceri, 1989).

As a final conclusion we can claim that we succeeded in modifying a known robust rank test (Fligner & Policello, 1981) in such a way that for small and moderate samples this variant (FPW), became a definitely better test of stochastic equality than any of its several possible alternatives in a broad range of different distributions. A similar statement can be formulated with respect to FPCW for moderate sample sizes, where it offers a good alternative to FPW. We note that the FPW and rW tests along with an interval estimation procedure for the A measure of stochastic superiority, are now available in the latest version of the Ministat Statistical Program Package (Vargha & Czigler, 1999).

With MiniStat we could also demonstrate that in practice one can really encounter situations where the comparison of means and stochastic comparison yield inconsistent statistical results. As an example see in Figure 1 the empirical distribution of the Rorschach variable Anat%[1] in a sample of low educated (n_L = 103), and in an independent sample of high educated (n_H = 114) persons.

Figure 1

The empirical distribution of Anat% in the groups of low educated (n = 103) and high educated (n = 114) persons. For these samples the means do not differ significantly (p > .50), whereas the null hypothesis of stochastic equality can be rejected (p < .05).



Comparing the two means ($\bar{x}_L$ = 8.2 and $\bar{x}_H$ = 8.3) via the Welch test the result was far from being significant ($W(163) = -.10$, $p > .50$), whereas FPW indicated a significant difference between the two samples ($\delta = .19$, $FPW(162) = -2.21$, $p < .05$). Since in this case

[1] Anat% is the percentage of Rorschach responses in a test protocol that contain an anatomic content.

$$Pr(\text{LowEd} > \text{HighEd}) = .39 \quad \text{and} \quad Pr(\text{LowEd} < \text{HighEd}) = .58,$$

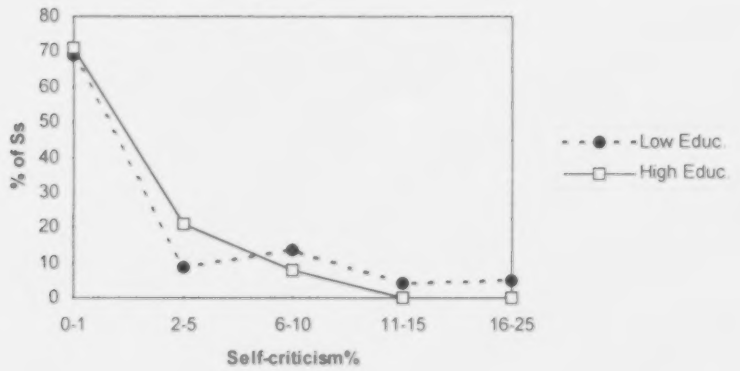
we can conclude that highly educated persons give relatively more anatomical responses in their Rorschach test than do undereducated persons.

Another example is shown in Figure 2. Here we compared the same two independent samples with respect to the Rorschach variable Self-criticism%[2]. In this case the relatively small number of very high scores in the LowEd sample caused the Welch test be significant ($\bar{x}_L = 2.7$, $\bar{x}_H = 1.2$, $W(138) = 2.81$, $p < .01$), while the small extent of stochastic difference between the two samples ($\delta = .08$) was not significant at all ($FPW(181) = .77$, $p > .40$). For this case

$$Pr(\text{LowEd} > \text{HighEd}) = .30 \quad \text{and} \quad Pr(\text{LowEd} < \text{HighEd}) = .22,$$

Figure 2

The empirical distribution of Self-criticism% in the groups of low educated (n = 103) and high educated (n = 114) persons. For these samples the means differ significantly (p < .01), whereas the null hypothesis of stochastic equality cannot be rejected (p > .40).



Finally we mention that the concept of stochastic equality can easily be generalized to more than two independent samples, and to the correlated samples case as well (see Vargha & Delaney, 2000).

References

Agresti, A. (1984). *Analysis of ordinal categorical data.* New York: Wiley.

Algina, J., Oshima, T. C., & Lin, W. Y. (1994). Type I error rates for Welch's test and James's second order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics, 19*, 275-291.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494-509.

[2] Self-criticism% is the percentage of Rorschach responses in a test protocol that can be characterized with the specific reaction self-criticism.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (rev. ed.).* New York: Academic Press.

Fligner, M. A., & Policello II, G. E. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association, 76,* 323-327.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks.* New York: Wiley.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361-365.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametris statistics. *The American Statistician, 35,* 124-129.

McKean, J. W., & Vidmar, T. J. (1994). A comparison of two rank-based methods for the analysis of linear models. *The American Statistician, 48,* 220-229.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

Onghena, P. (1993). A theoretical and empirical comparison of mainframe, microcomputer, and pocket calculator pseudorandom number generators. *Behavior Research Methods, Instruments, & Computers, 25,* 384-395.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., & Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics, 21,* 201-209.

Randles, R. H. & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics.* New York: Wiley.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Siegel, S., & Castellan, J. (1988). *Nonparametric statistics for the behavioral sciences (2nd ed.).* New York: McGraw-Hill.

Vargha, A., & Czigler, B. (1999). 3.2 verzió. A MiniStat statisztikai programcsomag, 3.2 verzió. [The MiniStat Statistical Program Package, version 3.2] Budapest: Pólya Kiadó.

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics, 23,* 170-192.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistic of McGraw and Wong. *Journal of Educational and Behavioral Statistics (in press).*

Wilcox, R. R. (1996). *Statistics for the social sciences.* San Diego, New York: Academic Press.

Zimmerman, D. W., & Zumbo, B. D. (1992). Parametric alternatives to the Student *t* test under violation of normality and homogeneity of variance. *Perceptual and Motor Skills, 74,* 835-844.

Zimmerman, D. W., & Zumbo, B. D. (1993a). Rank transformations and power of the Student *t* test and Welch t`test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47,* 523-539.

Zimmerman, D. W., & Zumbo, B. D. (1993b). The relative power of parametric and nonparametric statistical methods. In: G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 481-517). Hillsdale, NJ: Erlbaum.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology, 51,* 139-149.

Table 1

Skewness (α_3) and kurtosis (α_4) values of the lambda-distributions applied in the simulation.

Skewness	Kurtosis		
	Low	Moderate	High
symmetric	$\alpha_3 = 0$, $\alpha_4 = 1.8$	$\alpha_3 = 0$, $\alpha_4 = 3.0$	$\alpha_3 = 0$, $\alpha_4 = 9.0$
moderately asymmetric	$\alpha_3 = 1$, $\alpha_4 = 3.4$	$\alpha_3 = 1$, $\alpha_4 = 4.6$	$\alpha_3 = 1$, $\alpha_4 = 10.6$
heavily asymmetric	$\alpha_3 = 2$, $\alpha_4 = 8.6$	$\alpha_3 = 2$, $\alpha_4 = 9.8$	$\alpha_3 = 2$, $\alpha_4 = 15.8$

Table 2

Shift values ensuring stochastic equality for some couples of lambda distribution types depending on the ratio of standard deviations of the two distributions to be compared. The initially standardized X and Y lambda distributions were submitted to the $X' = \sigma_1 X$, $Y' = \sigma_2 Y + C$ linear transformations, where C is the corresponding tabled shift value.

Distribution		$\sigma_1:\sigma_2$						
Sample 1	Sample 2	4:1	3:1	2:1	1:1	1:2	1:3	1:4
$\alpha_3 = 0$ $\alpha_4 = 3$	$\alpha_3 = 0$ $\alpha_4 = 1.8$	0	0	0	0	0	0	0
$\alpha_3 = 0$ $\alpha_4 = 9$	$\alpha_3 = 1$ $\alpha_4 = 1.6$	.01	.02	.03	.05	.12	.21	.28
$\alpha_3 = 1$ $\alpha_4 = 4.6$	$\alpha_3 = 1$ $\alpha_4 = 4.6$	–.62	–.43	–.23	0	.23	.43	.62
$\alpha_3 = -1$ $\alpha_4 = 4.6$	$\alpha_3 = 1$ $\alpha_4 = 4.6$	.64	.46	.30	.17	.30	.47	.64
$\alpha_3 = 1$ $\alpha_4 = 4.6$	$\alpha_3 = 2$ $\alpha_4 = 8.6$	–.61	–.42	–.22	.04	.41	.77	1.11
$\alpha_3 = -1$ $\alpha_4 = 4.6$	$\alpha_3 = 2$ $\alpha_4 = 8.6$	.65	.49	.34	.24	.49	.78	1.11
$\alpha_3 = 2$ $\alpha_4 = 8.6$	$\alpha_3 = 2$ $\alpha_4 = 15.8$	–1.11	–.78	–.43	–.07	.18	.38	.58
$\alpha_3 = -2$ $\alpha_4 = 8.6$	$\alpha_3 = 2$ $\alpha_4 = 15.8$	1.13	.83	.52	.26	.35	.49	.64

Table 3

Summary statistics of empirical Type I error rates of six rank tests for testing the null hypothesis of stochastic equality against two-sided alternatives at 5% nominal level, with equal small sample sizes (m = n = 9). These statistics are based on 117 ($\sigma_1=\sigma_2$) or 234 ($\sigma_1\neq\sigma_2$) different pairs of distributions.

Mean of Type I error estimates (α = .05)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.052	.060	.069+	.077++
rW:	.052	.058	.063+	.066+
FP:	.051	.054	.057	.057
FPW:	.054	.055	.056	.055
FPC:	.059	.061+	.062+	.062+
FPCW:	.063+	.062+	.060	.058

Smallest Type I error rate (α = .05)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.050	.052	.059	.065
rW:	.049	.052	.057	.061
FP:	.050	.051	.054	.054
FPW:	.053	.053	.053	.051
FPC:	.058	.059	.059	.058
FPCW:	.061	.058	.055	.052

Largest Type I error rate (α = .05)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.057	.074	.086	.096
rW:	.056	.067	.069	.069
FP:	.054	.059	.059	.060
FPW:	.056	.058	.058	.058
FPC:	.061	.064	.064	.065
FPCW:	.065	.064	.063	.062

Table 3 (cont.)

Mean of Type I error estimates (α = .10)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.097	.106	.118	.127+
rW:	.097	.106	.118	.127+
FP:	.101	.103	.105	.108
FPW:	.097	.097	.098	.099
FPC:	.113	.116	.119	.123+
FPCW:	.109	.107	.105	.104

Smallest Type I error rate (α = .10)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.092	.096	.105	.115
rW:	.092	.096	.105	.115
FP:	.099	.099	.101	.102
FPW:	.095	.095	.096	.096
FPC:	.111	.112	.114	.116
FPCW:	.106	.102	.102	.101

Largest Type I error rate (α = .10)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.103	.122	.136	.145
rW:	.103	.122	.136	.145
FP:	.104	.108	.113	.119
FPW:	.101	.100	.102	.106
FPC:	.117	.121	.128	.135
FPCW:	.113	.111	.109	.108

Note: + Denotes mean estimates exceeding nominal level by more than 20%.
++ Denotes mean estimates exceeding nominal level by more than 40%.

Table 4

Summary statistics of empirical Type I error rates of six rank tests for testing the null hypothesis of stochastic equality against two-sided alternatives at 5% nominal level, with equal moderate sample sizes (m = n = 18). These statistics are based on 117 ($\sigma_1=\sigma_2$) or 234 ($\sigma_1\neq\sigma_2$) different pairs of distributions.

Mean of Type I error estimates (α = .05)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.053	.060	.069+	.076++
rW:	.053	.059	.067+	.072++
FP:	.060	.061+	.063+	.064+
FPW:	.051	.051	.051	.051
FPC:	.065+	.066+	.067+	.067+
FPCW:	.056	.055	.054	.054

Smallest Type I error rate (α = .05)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.050	.053	.060	.067
rW:	.050	.053	.059	.065
FP:	.058	.059	.061	.062
FPW:	.049	.049	.049	.049
FPC:	.063	.064	.065	.065
FPCW:	.054	.053	.052	.051

Largest Type I error rate (α = .05)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.059	.073	.083	.088
rW:	.059	.071	.078	.083
FP:	.062	.065	.067	.068
FPW:	.053	.053	.053	.053
FPC:	.067	.069	.069	.070
FPCW:	.058	.057	.056	.056

Table 4 (cont.)

Mean of Type I error estimates (α = .10)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.100	.110	.122+	.131+
rW:	.100	.110	.122+	.131+
FP:	.107	.109	.111	.113
FPW:	.097	.097	.097	.098
FPC:	.114	.115	.116	.116
FPCW:	.104	.103	.102	.101

Smallest Type I error rate (α = .10)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.095	.099	.110	.120
rW:	.095	.099	.110	.120
FP:	.103	.106	.108	.110
FPW:	.094	.095	.095	.095
FPC:	.111	.112	.113	.114
FPCW:	.102	.100	.099	.099

Largest Type I error rate (α = .10)

(σ_1:σ_2)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.108	.127	.139	.147
rW:	.108	.127	.139	.147
FP:	.110	.114	.116	.117
FPW:	.100	.100	.100	.101
FPC:	.117	.118	.119	.119
FPCW:	.107	.106	.104	.104

Note: + Denotes mean estimates exceeding nominal level by more than 20%.
++ Denotes mean estimates exceeding nominal level by more than 40%.

Table 5

Averaged empirical power rates of six rank tests for testing the null hypothesis of stochastic equality against the H_1: $A_{12} = .64$ alternative at 5% and 10% nominal levels. These statistics are based on 117 ($\sigma_1=\sigma_2$) or 234 ($\sigma_1\neq\sigma_2$) different pairs of distributions. Power rates whose corresponding mean Type I error rates exceeded the nominal level by more than 20% are in parentheses.

Power rates with α = .05 and m = n = 9					**Power rates with α = .10 and m = n = 9**				
(σ_1:σ_2)	**(1:1)**	**(1:2)**	**(1:3)**	**(1:4)**	**(σ_1:σ_2)**	**(1:1)**	**(1:2)**	**(1:3)**	**(1:4)**
rt:	.17	.18	(.19)	(.20)	**rt:**	.26	.26	.27	(.28)
rW:	.17	.18	(.18)	(.18)	**rW:**	.26	.26	.27	(.28)
FP:	.17	.17	.16	.16	**FP:**	.26	.25	.25	.25
FPW:	.17	.17	.16	.15	**FPW:**	.25	.25	.24	.23
FPC:	.18	(.18)	(.17)	(.17)	**FPC:**	.28	.28	.27	(.27)
FPCW:	(.19)	(.18)	.17	.16	**FPCW:**	.27	.26	.25	.24
Power rates with α = .05 and m = n = 18					**Power rates with α = .10 and m = n = 18**				
(σ_1:σ_2)	**(1:1)**	**(1:2)**	**(1:3)**	**(1:4)**	**(σ_1:σ_2)**	**(1:1)**	**(1:2)**	**(1:3)**	**(1:4)**
rt:	.31	.31	(.32)	(.32)	**rt:**	.42	.42	(.42)	(.43)
rW:	.31	.31	(.31)	(.31)	**rW:**	.42	.42	(.42)	(.43)
FP:	(.32)	(.31)	(.30)	(.29)	**FP:**	.43	.42	.40	.39
FPW:	.30	.28	.27	.26	**FPW:**	.41	.39	.38	.36
FPC:	(.34)	(.32)	(.31)	(.30)	**FPC:**	.44	.43	.41	.40
FPCW:	.31	.29	.28	.26	**FPCW:**	.42	.40	.38	.37

Table 6

Averaged empirical power rates of six rank tests for testing the null hypothesis of stochastic equality against the H_1: $A_{12} = .71$ alternative at 5% and 10% nominal levels. These statistics are based on 117 ($\sigma_1=\sigma_2$) or 234 ($\sigma_1\neq\sigma_2$) different pairs of distributions. Power rates whose corresponding mean Type I error rates exceeded the nominal level by more than 20% are in parentheses.

Power rates with α = .05 and m = n = 9					**Power rates with α = .10 and m = n = 9**				
(σ_1:σ_2)	**(1:1)**	**(1:2)**	**(1:3)**	**(1:4)**	**(σ_1:σ_2)**	**(1:1)**	**(1:2)**	**(1:3)**	**(1:4)**
rt:	.33	.34	(.35)	(.36)	**rt:**	.45	(.45)	.45	(.46)
rW:	.33	.33	(.33)	(.32)	**rW:**	.45	.45	.45	(.46)
FP:	.32	.32	.31	.30	**FP:**	.45	.44	.42	.42
FPW:	.33	.32	.30	.29	**FPW:**	.44	.42	.41	.40
FPC:	.35	(.34)	(.32)	(.31)	**FPC:**	.47	(.46)	.45	(.45)
FPCW:	(.35)	(.33)	.31	.29	**FPCW:**	(.46)	(.44)	.42	.41
Power rates with α = .05 and m = n = 18					**Power rates with α = .10 and m = n = 18**				
(σ_1:σ_2)	**(1:1)**	**(1:2)**	**(1:3)**	**(1:4)**	**(σ_1:σ_2)**	**(1:1)**	**(1:2)**	**(1:3)**	**(1:4)**
rt:	.60	.60	(.59)	(.59)	**rt:**	.71	.71	(.70)	(.70)
rW:	.60	.60	(.59)	(.58)	**rW:**	.71	.71	(.70)	(.70)
FP:	(.62)	(.60)	(.57)	(.56)	**FP:**	.72	.70	.68	.66
FPW:	.59	.56	.53	.51	**FPW:**	.70	.68	.65	.64
FPC:	(.63)	(.61)	(.58)	(.56)	**FPC:**	.73	.71	.68	.67
FPCW:	.60	.57	.54	.52	**FPCW:**	.71	.69	.66	.64

Table 7

Summary statistics of empirical Type I error rates of six rank tests for testing the null hypothesis of stochastic equality against two-sided alternatives at 5% nominal level, with unequal small sample sizes (m = 6, n = 12). These statistics are based on 117 different pairs of distributions.

Mean of Type I error estimates (α = .05)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.131++	.114++	.090++	.055	.037	.034	.034
rW:	.063+	.064+	.064+	.059	.061+	.069+	.077++
FP:	.064+	.066+	.065+	.052	.040	.037	.037
FPW:	.056	.058	.058	.051	.046	.047	.048
FPC:	.061+	.065+	.066+	.056	.045	.042	.042
FPCW:	.057	.061+	.063+	.060	.055	.054	.054

Smallest Type I error rate (α = .05)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.112	.093	.070	.041	.033	.032	.032
rW:	.057	.060	.061	.055	.054	.057	.065
FP:	.058	.061	.059	.042	.036	.035	.035
FPW:	.049	.053	.056	.047	.043	.043	.045
FPC:	.052	.057	.063	.048	.041	.040	.040
FPCW:	.049	.053	.058	.056	.052	.051	.051

Largest Type I error rate (α = .05)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.160	.144	.120	.078	.045	.038	.038
rW:	.068	.069	.069	.066	.073	.085	.093
FP:	.069	.070	.069	.062	.045	.041	.040
FPW:	.062	.063	.062	.059	.051	.051	.050
FPC:	.070	.072	.070	.065	.051	.047	.044
FPCW:	.065	.068	.068	.064	.059	.057	.056

Table 7 (cont.)

Mean of Type I error estimates (α = .10)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.192++	.177++	.151++	.105	.079	.075	.076
rW:	.127+	.118	.110	.104	.111	.122+	.130+
FP:	.142++	.131+	.119	.101	.087	.084	.085
FPW:	.112	.105	.100	.092	.086	.086	.087
FPC:	.132+	.125+	.119	.110	.100	.096	.095
FPCW:	.081	.089	.098	.102	.102	.103	.105

Smallest Type I error rate (α = .10)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.177	.158	.126	.085	.072	.071	.071
rW:	.116	.110	.105	.101	.100	.107	.117
FP:	.130	.122	.110	.091	.082	.081	.081
FPW:	.104	.100	.096	.087	.082	.082	.083
FPC:	.122	.118	.114	.102	.095	.092	.092
FPCW:	.067	.075	.087	.097	.095	.097	.100

Largest Type I error rate (α = .10)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.211	.202	.183	.137	.090	.081	.084
rW:	.149	.136	.121	.110	.126	.136	.141
FP:	.164	.152	.135	.113	.095	.089	.089
FPW:	.132	.120	.107	.099	.092	.091	.092
FPC:	.151	.140	.127	.117	.106	.101	.098
FPCW:	.090	.099	.107	.108	.106	.108	.110

Note: + Denotes mean estimates exceeding nominal level by more than 20%.
++ Denotes mean estimates exceeding nominal level by more than 40%.

Table 8

Summary statistics of empirical Type I error rates of six rank tests for testing the null hypothesis of stochastic equality against two-sided alternatives at 5% nominal level, with unequal small sample sizes (m = 12, n = 24). These statistics are based on 117 different pairs of distributions.

Mean of Type I error estimates (α = .05)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.116++	.103++	.083++	.051	.034	.032	.032
rW:	.066+	.063+	.058	.054	.060	.071++	.080++
FP:	.078++	.077++	.074++	.065+	.058	.056	.056
FPW:	.055	.054	.052	.048	.046	.047	.047
FPC:	.076++	.076++	.075++	.070+	.065+	.063+	.063+
FPCW:	.054	.053	.053	.052	.052	.053	.054

Smallest Type I error rate (α = .05)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.103	.087	.065	.037	.030	.029	.030
rW:	.062	.059	.054	.050	.053	.060	.069
FP:	.075	.074	.070	.060	.055	.054	.054
FPW:	.053	.052	.049	.045	.043	.044	.045
FPC:	.072	.074	.072	.067	.062	.062	.061
FPCW:	.051	.052	.051	.050	.049	.051	.052

Largest Type I error rate (α = .05)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.135	.125	.108	.072	.041	.035	.036
rW:	.069	.069	.065	.058	.074	.087	.096
FP:	.080	.079	.079	.072	.062	.059	.058
FPW:	.057	.056	.055	.051	.048	.049	.049
FPC:	.078	.078	.077	.074	.067	.066	.066
FPCW:	.056	.055	.054	.053	.054	.055	.055

Table 8 (cont.)

Mean of Type I error estimates (α = .10)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.184++	.170++	.145++	.100	.074	.071	.072
rW:	.120	.115	.109	.103	.114	.129+	.141++
FP:	.129+	.126+	.122+	.113	.104	.102	.102
FPW:	.102	.101	.098	.093	.091	.092	.093
FPC:	.126+	.125+	.123+	.119	.114	.113	.112
FPCW:	.099	.099	.099	.099	.100	.102	.102

Smallest Type I error rate (α = .10)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.169	.151	.121	.081	.068	.067	.068
rW:	.114	.108	.103	.098	.102	.114	.127
FP:	.124	.122	.117	.107	.101	.100	.100
FPW:	.099	.097	.094	.090	.088	.089	.090
FPC:	.122	.121	.119	.115	.111	.110	.110
FPCW:	.097	.096	.096	.096	.097	.099	.099

Largest Type I error rate (α = .10)

(σ_1:σ_2)	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	.203	.196	.176	.129	.085	.077	.076
rW:	.132	.126	.118	.110	.136	.153	.163
FP:	.137	.133	.130	.120	.110	.106	.105
FPW:	.110	.106	.102	.097	.095	.095	.095
FPC:	.132	.129	.127	.123	.118	.115	.115
FPCW:	.105	.102	.101	.101	.103	.104	.105

Note: + Denotes mean estimates exceeding nominal level by more than 20%.
++ Denotes mean estimates exceeding nominal level by more than 40%.

Table 9

Averaged empirical power rates of six rank tests for testing the null hypothesis of stochastic equality against the H_1: $A_{12} = .64$ alternative at 5% and 10% nominal levels. These statistics are based on 117 different pairs of distributions. Power rates whose corresponding mean Type I error rates exceeded the nominal level by more than 20% are in parentheses.

Power rates with $\alpha = .05$, m = 6 and n = 12

$(\sigma_1:\sigma_2)$	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	(.24)	(.22)	(.19)	.16	.14	.13	.14
rW:	(.13)	(.14)	(.15)	.17	(.20)	(.22)	(.24)
FP:	(.13)	(.14)	(.15)	.15	.14	.14	.14
FPW:	.12	.13	.14	.15	.16	.17	.17
FPC:	(.13)	(.14)	(.16)	.16	.16	.15	.15
FPCW:	.12	(.14)	(.15)	.17	.18	.18	.18

Power rates with $\alpha = .1$, m = 6 and n = 12

$(\sigma_1:\sigma_2)$	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	(.32)	(.30)	(.28)	.25	.23	.23	.23
rW:	.23	.22	.23	.25	.29	(.31)	(.32)
FP:	(.25)	(.24)	.24	.24	.24	.24	.24
FPW:	.21	.20	.21	.23	.24	.25	.25
FPC:	(.24)	(.23)	.24	.26	.27	.26	.26
FPCW:	.16	.18	.20	.25	.27	.28	.28

Power rates with $\alpha = .05$, m = 12 and n = 24

$(\sigma_1:\sigma_2)$	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	(.32)	(.31)	(.29)	.27	.25	.25	.25
rW:	(.22)	(.23)	.24	.29	.35	(.38)	(.40)
FP:	(.25)	(.26)	(.28)	(.31)	.33	.33	.33
FPW:	.20	.21	.23	.27	.30	.30	.30
FPC:	(.24)	(.26)	(.28)	(.32)	(.35)	(.35)	(.34)
FPCW:	.20	.21	.23	.28	.32	.32	.32

Power rates with $\alpha = .1$, m = 12 and n = 24

$(\sigma_1:\sigma_2)$	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	(.41)	(.41)	(.40)	.38	.37	.37	.37
rW:	(.33)	.33	.35	.40	.46	(.49)	(.51)
FP:	(.34)	(.35)	(.37)	.41	.44	.44	.43
FPW:	.29	.31	.33	.38	.41	.42	.41
FPC:	(.33)	(.35)	(.37)	.42	.45	.46	.45
FPCW:	.29	.30	.33	.39	.43	.44	.43

Table 10

Averaged empirical power rates of six rank tests for testing the null hypothesis of stochastic equality against the H_1: $A_{12} = .71$ alternative at 5% and 10% nominal levels. These statistics are based on 117 different pairs of distributions. Power rates whose corresponding mean Type I error rates exceeded the nominal level by more than 20% are in parentheses.

Power rates with $\alpha = .05$, m = 6 and n = 12

$(\sigma_1:\sigma_2)$	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	(.37)	(.35)	(.33)	.30	.29	.28	.28
rW:	(.23)	(.24)	(.27)	.32	(.38)	(.41)	(.43)
FP:	(.23)	(.25)	(.27)	.30	.30	.29	.29
FPW:	.21	.23	.26	.30	.33	.33	.33
FPC:	(.22)	(.25)	(.28)	.31	.32	.31	.31
FPCW:	.22	(.24)	(.27)	.33	.35	.35	.35

Power rates with $\alpha = .1$, m = 6 and n = 12

$(\sigma_1:\sigma_2)$	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	(.46)	(.45)	(.44)	.43	.42	.42	.42
rW:	.36	.36	.37	.43	.50	(.52)	(.53)
FP:	(.39)	(.38)	.39	.42	.44	.44	.44
FPW:	.33	.33	.35	.41	.44	.44	.45
FPC:	(.37)	(.37)	.39	.44	.46	.46	.46
FPCW:	.26	.29	.34	.43	.47	.48	.48

Power rates with $\alpha = .05$, m = 12 and n = 24

$(\sigma_1:\sigma_2)$	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	(.54)	(.54)	(.54)	.55	.54	.54	.54
rW:	(.42)	(.44)	.48	.57	.66	(.69)	(.70)
FP:	(.46)	(.48)	(.53)	(.60)	.64	.63	.63
FPW:	.40	.42	.46	.54	.60	.61	.60
FPC:	(.45)	(.48)	(.53)	(.61)	(.66)	(.65)	(.65)
FPCW:	.39	.42	.46	.56	.62	.63	.62

Power rates with $\alpha = .1$, m = 12 and n = 24

$(\sigma_1:\sigma_2)$	(4:1)	(3:1)	(2:1)	(1:1)	(1:2)	(1:3)	(1:4)
rt:	(.64)	(.65)	(.66)	.67	.68	.68	.68
rW:	(.55)	.57	.60	.69	.76	(.79)	(.79)
FP:	(.57)	(.59)	(.63)	.70	.74	.74	.73
FPW:	.52	.54	.58	.67	.72	.72	.71
FPC:	(.56)	(.58)	(.63)	.71	.75	.75	.75
FPCW:	.51	.54	.58	.68	.73	.74	.73